The Annals of Statistics
2002, Vol. 30, No. 6, 1649–1665

JOHN W. TUKEY'S CONTRIBUTIONS TO ANALYSIS OF VARIANCE

BY T. P. SPEED

University of California, Berkeley

John Tukey connected the theory underlying simple random sampling without replacement, cumulants, expected mean squares and spectrum analysis. He gave us one degree of freedom for nonadditivity, and he pioneered finite population models for understanding ANOVA. He wrote widely on the nature and purpose of ANOVA, and he illustrated his approach. In this appreciation of Tukey's work on ANOVA we summarize and comment on his contributions, and refer to some relevant recent literature.

1. Introduction. Most (9/15) of John Tukey's contributions to analysis of variance (hereafter ANOVA) can be found in Volume 7 of *The Collected Works of John W. Tukey* [17]. Also in that volume are two items which will be of interest to readers of this paper. One is a six-page foreword to the nine collected papers by John Tukey himself. The other is an historical introduction to and remarks on the roles of analysis of variance, and some brief comments on the individual papers by the volume editor, David R. Cox. However, Tukey being Tukey, there is no substitute for reading the papers themselves. Every one of them advances our knowledge, at times dramatically, while seeming to be no more than a lucid exposition from first principles of some well-established part of our subject. There are exceptions to this last statement.

John Tukey's main published contributions to ANOVA were made in a little over a decade, from 1949 to 1961. They constitute approximately 20% of his output over this period, and so about 5% of his total output. In subject matter these papers range from the foundational to the computational, from the algebraic to the interpretational, and contain some strikingly original views of the topics he discusses. How many of us see a clear connection between finite-population simple random sampling as in books on sampling, Fisher's k-statistics and cumulants for calculating moments of sample moments, the moments of mean squares in ANOVA tables and the arithmetic of spectrum analysis? At the same time as he was clarifying the analysis of variance qua variance, he highlighted the importance of scale to the notion of interaction in the analysis of means, and gave us a tool for identifying and removing removable nonadditivity. He also showed us how to analyze a complex multifactorial data set; indeed in no fewer than four of the

Received January 2002; revised March 2002.

AMS 2000 subject classifications. Primary 62J10; secondary 94A20.

Key words and phrases. Odoffna, ANOVA, moments, cumulants, k-statistics, polykays, variances, components of variance, mean squares, factorials, interactions, pigeonhole model.

papers below we get his views on the nature and purpose of ANOVA. It was much broader than the usual one which focusses on testing.

In my opinion much of Tukey's work on ANOVA is underappreciated, and much of that which was appreciated at the time has been forgotten. He laments [17, page lii], wrongly as it turns out, "Perhaps regrettably, I am not aware of very much that extends papers 5, 6, 7, and 9" (of [17]). Some of his work on ANOVA, for example, his "dyadic ANOVA" and his "components in regression," was never followed up. Neither of these titles scores a hit (with Tukey's meaning) in *Current Index to Statistics*. Fashions change, and the foundational worries or solutions of one generation of statisticians can cease to be of interest to a later generation. It is for this reason as well as to celebrate Tukey's genius that it is a real pleasure to be able to remind readers of his wonderful contributions to ANOVA, including creating the abbreviation itself.

2. ODOFFNA. How we will miss Tukey's neologisms. His one degree of freedom for nonadditivity (ODOFFNA) paper [2] is perhaps his best-known and most striking contribution to the analysis of variance and needs little introduction here. Whereas others had paid attention to nonconstancy of the variance or nonnormality of the "errors" in ANOVA, Tukey was concerned with nonadditivity. Explaining his ideas in the context of a singly replicated row-by-column table, he showed how to isolate a single degree of freedom from the "residual," "error" or "interaction" sum of squares ("call it what you will" he said), and so test the null hypothesis of additivity using a statistic which gave power against a restricted class of multiplicative alternatives. The statistic was motivated by the idea of a power transformation; it was illustrated graphically through three examples, and some elegant distribution theory was presented. This is a gem of a paper and amply deserves its place in the texts [see, e.g., Scheffé (1959) or Seber (1977)]. Tukey's later papers [5, 13] on the same topic present no new ideas; rather they illustrate the earlier ideas in more general contexts, something he pointed out was possible in [2].

What has happened to ODOFFNA since the 1960s? These days most people concerned about the possibility that their linear model might better satisfy the standard assumptions of additivity, homoscedasticity and normality of errors after a transformation will make use of the Box and Cox (1964) theory. However, their approach to transformations is not a complete substitute for ODOFFNA, as can be seen in Tukey's [14] discussion of additive and multiplicative fits to two-way tables (see especially [14], Section 10F). It is likely that we will continue to extract ODOFFNA in new contexts in the future, and for more on this, see Tukey's own comments on the follow-up to ODOFFNA in his foreword to [17].

3. Complex analyses of variance: general problems [11]. In [11] Green and Tukey made a number of general points concerning complex analyses of variance in the course of analyzing a specific experimental data set. Some of the points are familiar, some were new at the time but most are still of interest today. The authors explain that the purpose of their analysis is “to provide a simple summary of the variation in the experimental data, and to indicate the stability of means and other meaningful quantities extracted from the data.” They intended their approach to be in opposition to the view that the sole purpose of ANOVA is to provide tests of significance. It was aimed at researchers in psychology and followed a review of the use of ANOVA in that field a few years earlier.

The experiment is from psychophysics and involves six factors: sex (S, two levels), sight (I, two levels), persons (P, eight levels), rate (R, four levels), weight (W, seven levels) and date (D, two levels). All of S, I, R, W and D are crossed, while P is nested in a balanced way within $S \times I$ so we may describe the factor relationships by the formula $((S \times I)/P) \times R \times W \times D$. The response was a difference limen, a kind of threshold of perception, which could be expressed as a difference in weights, a squared difference in weights, a ratio of weights, a logarithm of a ratio of weights or even a response time.

One novel aspect of this paper is that the authors discuss not only what scale to use for the dependent variable; that is, possible transformations, but also just what the dependent variable should be in that context: a difference, a squared difference, a ratio, a log ratio, etc. After an initial analysis with one response variable, they choose another and obtain a new, and to their minds better, analysis. Another novelty at that time was the careful discussion of the nesting and crossing between factors and their implications for the analysis. This was no doubt inspired by the discussion of these matters Tukey and Cornfield gave in [8], which was published some four years before [11].

Perhaps the most interesting part of this paper is the extended section “Variance components and the proper error term” and the section “Variance components in the illustrative example” which follows it. The first of these discusses an example simpler than the actual experiment and draws heavily on material concerning the pigeonhole model in [8]; see Section 4.3. Then they turn to the experiment and things get interesting when they seek to impose a sampling model on the factors. The four levels of rate (50, 100, 150 and 200 g/s) and the seven levels of weight ($100, 150, \dots, 400$ g) are admitted to present a problem for their pigeonhole model. Are they exhaustive samples from finite populations, that is, fixed; are they small samples from large populations of levels, that is, random; or are they something else? Whereas it was easy for them to view sex and sight (blind or not) as fixed, and person as random, the choice for R and W was far less obvious. After some discussion of various options, including a mention of using polynomials to fit responses to rate and weight, they decide to regard R and W as random “although we recommend against this procedure [for scaled variables] in general.” The ideal that one ANOVA theory fits all cases seems hard to live up to, even when you are the creator of the theory.

As soon as all factors are assigned the category fixed or random, it is possible to write out all 39 expected mean square lines of the ANOVA table, and this they do. Next follows an illuminating discussion of "aggregation and pooling" of lines in the table, which, when implemented with the illustrative data, reduces the 39 lines to 15. They make use of a modified version of a procedure of Paull (1950) which Tukey highlights in his Introduction to [17] and seems to be of interest today. There are two useful graphical representations of the relative contributions of the different sources of variability, one in two dimensions which is especially appealing, but on the whole there is relatively little plotting of the data, a large contrast with Tukey's later work, for example, in [14].

A later analysis of this same data set can be found in Johnson and Tukey [15]. Looking back on this paper after four decades, and bearing in mind all that Tukey wrote on ANOVA before and after that time, one cannot help but be struck by how little use he made in this paper of the processes and procedures he recommended when considering such an analysis. Referring to matters to be discussed in Section 4, he made no attempt to assign standard errors to his estimated variance components, under either normality or any other assumptions, the scientific purpose of the experiment was nowhere mentioned, the situations or populations to which inference was to be made were nowhere mentioned, even the means he calculated and plotted were not assigned any measures of their stability, something that was stated at the beginning of the paper to be one of the major purposes of ANOVA. Granted this was an expository paper with a limited objective, and probably already long by the standards of the journal, but I think the point remains that it is hard to put Tukey's ANOVA theory into practice, even for Tukey himself.

4. Some moment calculations.

Tukey wanted to derive average values and variances and later a third moment of consider later. He tells us [17, page liv] that his first attempt at deriving the variance of the between variance component in an unbalanced one-way design took five or six full days "using old-fashioned clumsy methods." He was "convinced that it ought not to be so hard" and so "went looking for better tools, and eventually came out with the polykays." Polykays are generalizations of Fisher's k-statistics and we now outline the main points from the papers in which they were introduced.

4.1. *Some sampling simplified*; *keeping moment-like computations simple* [3, 6].

In 1929 Fisher introduced k-statistics as unbiased estimators of cumulants and a computational technique which radically simplified much previous research on moments of moments. It would take us too far astray to describe his technique in detail [see Speed (1986a)], but we can describe the simplest of his results in this area as soon as we recall the following well-known facts. If $X_1, \dots, X_n$ are i.i.d. random variables with common first two cumulants κ_1 and κ_2 (the mean and variance, respectively), then

$$k_1 = \frac{1}{n}\sum X_i = \bar{X} \quad \text{and} \quad k_2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$

satisfy

$$\mathbf{E}k_1 = \kappa_1 \quad \text{and} \quad \mathbf{E}k_2 = \kappa_2.$$

Now the k's are Fisher's k-statistics, that is, unbiased estimates of the corresponding cumulants. The key results of Fisher (1929) were the general definition of k-statistics and a procedure for calculating their joint cumulants whose core was a rule for calculating the coefficients of lower order k-statistics in an expansion for the product of two k-statistics. The relationships above are the simplest relevant results: the expected values or first cumulants of the first two k-statistics. Next would come the results which come from replacing $\mathbf{E}$ by var or covar; that is, replacing first by second cumulant in the sample-population calculation.

We all know that $\text{var}(k_1) = \kappa_2/n$, but what about $\text{var}(k_2)$? This result, first derived by Gauss, is not quite so well known, but turns out to be

$$\text{var}(k_2) = \frac{2}{n-1}\kappa_2^2 + \frac{1}{n}\kappa_4.$$

Deriving this last fact is already messy enough to warrant thinking very carefully about the algebraic formulation one adopts, and any desire to obtain more general expressions of the same kind focusses the mind greatly on the same issue. Fisher had his approach, Tukey simplified it as we shall see and it can be simplified yet again; see Speed (1983) and McCullagh (1987).

Tukey's main aim in [3] and [6] was to extend these results (and others like them) to the finite population case. Apparently unknown to Tukey, this task had been begun by Neyman in 1923 [see Neyman (1925)], though far less elegantly or generally. To achieve his aim Tukey extended Fisher's entire machinery. He named the tool he developed polykays—multiply-indexed generalizations of k-statistics—later noting that these same functions had been introduced earlier by Dressel (1940) in a paper that was not noticed at the time. For Tukey polykays of order or weight r are indexed by partitions of the natural number r. For example, there are two of order 2, indexed by $(1, 1)$ and (2); three of order 3, indexed by $(1, 1, 1)$, $(1, 2)$ and (3); four of order 4, indexed by $(1, 1, 1, 1)$, $(1, 1, 2)$, $(1, 3)$ and (4); and so on. Fisher's k-statistics are the single subscript versions of the polykays, (1), (2), (3), (4) etc., hence Tukey's name. In what follows we drop the commas and parentheses in the partition notation, writing 1, 11, 2, etc.

How are polykays defined in general? To do this Tukey made use of an auxiliary class of symmetric functions also labelled by partitions, which he called symmetric means or, more simply, brackets, denoted by $\langle 1\rangle$, $\langle 11\rangle$, $\langle 2\rangle$, etc. These functions had the appealing property of rather transparently being "inherited on the average," which means that the average of the sample function over simple random sampling without replacement from a finite population was just the corresponding population function. Tukey avoided using the term "unbiased" as (so he said) "there are now so many kinds of unbiasedness!" The sample mean

$$\langle 1\rangle = \frac{1}{n}\sum x_i$$

is clearly inherited on the average, as is

$$\langle 11\rangle = \frac{1}{n(n-1)}\sum_{i\neq j} x_i x_j.$$

The value of brackets lies in the fact that [3, page 111] "every expression which is (i) a polynomial, (ii) symmetric, (iii) inherited in the average, can be written as a linear combination of brackets with coefficients which do not depend on the size of the set of numbers involved." As one illustration we give the following well-known and useful representation:

$$\frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{n}\sum x_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j} x_i x_j,$$

where the last two terms are transparently inherited in the average, neatly proving that the first term is also, a standard fact from sampling theory. Tukey would write this last relationship $(2) = \langle 2\rangle - \langle 11\rangle$, and in general he needed a rule giving the coefficients of brackets in the expansion of his parentheses (polykays). As he said ([3], page 124) "the single-index brackets have the coefficients for moments in terms of cumulants (given numerically by Kendall [(1943), Section 3.13] up to the 10th moment). The coefficients of brackets with several indices can be found by formal multiplication."

How do we use all this machinery? Elegant though it is, there is still some hard work: the multiplication tables need to be derived. Tukey derived his own, but by the time of publication of [3, 6] comprehensive tables had independently appeared [Wishart (1952a, b)]. A simple instance of a multiplication rule is

$$(*) \qquad (2)^2 = (22) + \frac{1}{n}(4) + \frac{2}{n-1}(22)$$

Let us see how this leads very painlessly to the main result of Neyman (1925). First, note that the preceding identity has a version connecting population k-statistics which is of the same form, but with n replaced by N. Next recall that the polykays (22), (4), etc. are all "inherited on the average." We now take the expectation (i.e., average) of $(*)$ over all samples and subtract from the result the population version of $(*)$. This leaves us with

$$\operatorname{var}((2)) = \left[\frac{1}{n} - \frac{1}{N}\right](4) + 2\left[\frac{1}{n-1} - \frac{1}{N-1}\right](22),$$

which is the formula Neyman worked hard to obtain. This was indeed "sampling simplified." Note also that if we let $N \to \infty$ (so-called infinite population) and use the easily proved fact that, in this case, (22) is just $(2)^2$, we obtain Gauss' result.

Tukey certainly simplified sampling. He demonstrated clearly that indeed finite populations are simpler to deal with, and more powerful, and he now had the machinery to carry out certain calculations in ANOVA.

Later developments cast Tukey's work in the framework of tensors [cf. Kaplan (1952) and, most recently within the general theory of symmetric functions, Speed (1986a)]. The gains from so doing are modest, but I think definitely worthwhile. One consequence of the tensor formulation is that some of Tukey's formal calculations (e.g., his symbolic o-multiplication) cease to be "tricks." Another is the greater simplicity which comes from allowing all random variables to be potentially different. For example, instead of calculating variances of variances, we calculate covariances of distinct covariances, and obtain variances by appropriately equating arguments. With this slightly greater generality, $(*)$ above becomes [Speed (1986a), page 43]

$$(12) \otimes (34) = (12|34) + \frac{1}{n}(1234) + \frac{1}{n-1}[(13|24) + (14|23)],$$

where 1, 2, 3 and 4 all label distinct variables. This simplification removes certain multiplicity factors and then reveals the coefficients defining polykays to be values of the Möbius function over a partition lattice, which I think is a real step forward; see Speed (1983) and McCullagh (1987).

Where are polykays now? There was a little theoretical development of them after Tukey's work, but he left no major problems unaddressed. I extended them to multiply-indexed arrays in Speed (1986a, b) and Speed and Silcock (1988a), and used the extensions to generalize the calculations of Tukey discussed in the next section. Apart from my own work the most recent references to polykays are Tracy (1973) and, an application of them, McCullagh and Pregibon (1987). To my knowledge there have been no other publications concerning polykays since then. In short, it seems that after about 25 years of life, polykays have been dead or sleeping for 25 years. Apparently they have served their purpose, though I have no doubt that they will be resurrected, awakened or reborn at some time in the future, when another problem comes along for whose solution they are the natural tool.

4.2. *Variances of variance components* [7, 9, 10]. Why did Tukey go to all the trouble of inventing polykays and their calculus, and what did he learn from so doing? Giving as one purpose of the analysis of variance "to estimate the sizes of the various components contributed to the overall variance from the corresponding sources," he wanted "to obtain formulas for the variances of the natural estimates of these variance components." Along with Gauss, Fisher and many others, Tukey wanted to go beyond normality, but almost uniquely he did so in dispensing with infinite populations. He regretted ([7], page 157) that he still had to leave "the customary (and dangerous) independence assumptions" concerning the terms in his linear population models. This answers the question "Why?" Let us now see some of what he learned in a simple case: the balanced single (or one-way) classification. Tukey's model for this takes the form

$$x_{ij} = \mu + \eta_i + \omega_{ij}, \qquad i = 1, \dots, c,\ j = 1, \dots, r,$$

where the $\{\eta_i\}$ are sampled from a population of size n with k-statistics $k_1, k_2, \dots,$ the $\{\omega_{ij}\}$ are from a population of size N with k-statistics $K_1, K_2, \dots$ and the samplings are independent and order randomized. If we denote by B and W the usual between-class and within-class mean squares, respectively, with expectations k_2 and K_2, then Tukey showed, among other results, that

$$\begin{aligned}\operatorname{var}(B) &= \left(\frac{1}{c} - \frac{1}{n}\right)k_4 + 2\left[\frac{1}{c-1} - \frac{1}{n-1}\right]k_{22}\\ &\qquad + \frac{4}{r(c-1)}k_2K_2 + \frac{2(rc-1)}{r^2c(r-1)(c-1)}K_{22},\\ \operatorname{cov}(B, W) &= -\frac{2}{rc(r-1)}K_{22},\\ \operatorname{var}(W) &= \left[\frac{1}{rc} - \frac{1}{N}\right]K_4 + 2\left[\frac{1}{c(r-1)} - \frac{1}{N-1}\right]K_{22}.\end{aligned}$$

The remainder of [7] consists of more formulae of this kind, derived for other variance component models: row-by-column classifications, Latin squares, balanced incomplete blocks and more general balanced models.

Paper [9] considers the special, more complicated case of an unbalanced one-way classification. One novelty here is that there is no single compelling estimate of the between-class component of variance, and so Tukey considers a class of estimates involving weights which need to be specified. He then derives the variances and covariances as before, generalizing those just given, and presents numerical examples. Lastly, paper [10] does what its title says: it presents the third moment about the mean, that is, the third cumulant of the quantity W given above.

What can we learn from or do with such formulae? In the first place, we can obtain qualitative insights by comparing the general finite population results with the special case of infinite normal populations. There k_4 and K_4 vanish, while k_{22} and K_{22} are k_2^2 and K_2^2, respectively, and of course $N = \infty$. In this case the results are familiar, and the extent to which the normal variances for the estimated variance components are too small or too large could, in principle, be examined. Interestingly, Tukey does not present formulae giving unbiased estimates of either the individual terms in his expressions for the variances of the estimated variance components, or for the variance expression as a whole. I would be very surprised if he did not have such formulae, for example, for k_4 and K_4 and k_{22} and K_{22} above, but he makes no mention of them. Without them, his aim of calculating estimates of the precision of estimated variance components under these more general assumptions must remain unfulfilled.

What has been done since the 1950s in this area? There has been more work on the topic of variances of estimated components and variance; see, for example, Harville (1969), but there, as in all other such cases that I know, the calculations are carried out under an assumption of normality. In some of my own work [Speed

(1986a, b), Speed and Silcock (1988a, b)] I have tried to extend Tukey's work to ANOVA models which are not built up additively from independent components.

4.3. *Average values of mean squares in factorials* [8]. This is an interesting and important paper: broad in coverage, profound in its analysis, beautifully written and elegant in its dealing with messy algebraic details. It is arguably Tukey's most important contribution to ANOVA. By the early to mid-1950s it was becoming clear that the concise description in Eisenhart (1947) of models for ANOVA did not provide a foundation for all uses of ANOVA. The now well-known mixed-model ambiguity concerning the interaction component of variance when (say) rows are "fixed" and columns "random" had emerged: in some linear model formulations this component appeared in the expected mean square line for both rows and columns, while in others it did not. It was apparent to many that the combining of linear models and ANOVA was not as simple as might have seemed at first. Neyman and his Polish colleagues found this out the hard way in 1935, but made no later attempt at a broad synthesis. Kempthorne (1952) in Ames, building on the work of Neyman and co-workers, Fairfield Smith in Raleigh, Tukey in Princeton, Cornfield at the National Institutes of Health in Bethesda and no doubt others elsewhere all sought to devise models of differing breadth and flexibility which would specialize appropriately under different assumptions, and lead to the desired analyses and inferences. Throughout all this, Fisher was silent on the topic, apparently holding to his view that "the analysis of variance is . . . a convenient method of arranging the arithmetic."

Anyone who reads the five sections comprising the Initial discussion of [8] quickly realizes that providing a general framework for ANOVA is no mean task. The subsequent six sections spelling out Cornfield and Tukey's approach prior to their presenting any average values shows that theirs is not an easy resolution. So it should come as no surprise when I say that the situation today is hardly any better than it was then in the mid-1950s. Cornfield and Tukey's paper should be essential reading for all those who care about these matters. But it is not read, and neither their approach nor any other has taken root among the legions of users of ANOVA and linear models. No treatment of the issues that prompted them to write that paper has yet gained acceptance; see below.

What are the issues? Although in most of his writings on ANOVA Tukey emphasized estimation of variance components above significance testing, this paper is very much motivated by testing. Expressions for average values of mean squares in factorials are the primary basis for testing: they tell us which mean squares can usefully be compared with which; that is, they dictate the choice of error term. So attention focuses sharply on the model assumptions leading to these averages. As Tukey and Cornfield point out in Section 2 of their paper, the choice among assumptions is important and is not simple. It includes but goes beyond empirical questions about the behavior of the experimental material. Assumptions must also depend on the nature of the sampling and randomization involved in obtaining the data, and the purpose of the analysis, as expressed by the situations or populations to which one wishes to make statistical inference.

Cornfield and Tukey's way ahead is by the use of what they call a pigeonhole model, in which combinations of experimental factors (rows, columns, etc.) define pigeonholes containing a finite or infinite population. If, like them, we illustrate ideas with the replicated row-by-column classification, then their assumption is that a sample of r rows is drawn from a possible R, and a sample of c columns is drawn from a possible C. These rc intersections define the pigeonholes which are the cells of the actual experiment, and from each of the rc cells a sample of n elements is drawn. "All the samplings—of rows, of columns, and within pigeonholes—are at random and independent of one another." This is their approach. They discuss at considerable length the way in which an equivalent linear model can be defined, making it clear just how different their linear model was from those previously used (and used today). Of particular significance was their notion of "tied" interaction, their avoidance of what they term the "special and dangerous" assumption of independence of the variation of interaction terms of main effects terms.

After their lengthy preliminaries it is almost a relief to get to the algebraic part of the paper: definitions of components of variance and rules for calculating what we now term expected mean squares. They discuss two-way and three-way designs in detail and give rules for designs with factors nested or crossed in arbitrary ways. There is an interesting discussion of the nature of the various proofs then extant of the formulae. At that time there were two types: "Proofs using special machinery or indirect methods (e.g., symmetry arguments and equating of coefficients for special assumptions)," the approach preferred by Tukey, and "proofs using relatively straightforward algebra," which was the preferred way of Cornfield. Neither of these was particularly effective in full generality.

The mathematical content of [8] has been revisited at least twice since 1956. The first time was by Haberman (1975), in a dense paper which does not seem to have been widely read. He makes effective use of the calculus of tensor products of vector spaces to give very concise proofs of the main results. A quite different approach was used in Speed (1985) [see also Speed and Bailey (1987)] (also not widely read), where the discussion was expressed in terms of the eigenvalues of the associated dispersion matrices. Other, less general formulations can be found in books on linear models and ANOVA, for example, Searle, Casella and McCulloch (1992).

As suggested earlier, all attempts at providing a general framework for ANOVA since 1956 should have come to terms with the material in [8]: they should either incorporate it or suggest an alternative approach. There have been many such attempts over the last 45 years, with Nelder (1977) providing the most far-reaching alternative, building on Nelder (1965a, b). This paper and especially the discussion of it are well worth reading, especially today. The most recent discussions of the "mixed models controversy" [see, e.g., Schwarz (1993) and Voss (1999) and references therein] refer to neither Cornfield and Tukey, Kempthorne, Nelder nor any other of the earlier generation of researchers in this area. Plackett (1960) gives an excellent review of this early work.

Tukey's contribution to the discussion of Nelder (1977) is particularly interesting, in part because it reveals so clearly his distrust in models. It should be read in full, but here are some tantalizing excerpts, all the more relevant when one bears in mind that all recent discussion of this issue is a discussion of models:

> I join with the speaker in hoping for an eventual and agreed-upon description. I hope the present paper will help us approach this ideal state, but I must say that it has not brought us there.
>
> Three types of variability arise in almost any question about a set of comparative measurements, experimental or not: measurement variability, sampling variability and contextual variability.
>
> A major point, on which I cannot yet hope for universal agreement, is that our focus must be on questions, not models.
>
> One conclusion I draw from such examples is this: Models can—and will—get us into deep trouble if we expect them to tell us what the unique proper questions are.

I close this section with some personal comments, but before I do so, I should confess that I too have attempted to publish a description of ANOVA which I had hoped might have become "agreed-upon." It did not even get accepted for publication. However, I think I represent more than myself when I say that, for all my admiration of [8] and what it attempted to do, that solution was simply too far away from the world of linear models most of us inhabit. In my view, and I suspect that of many others, linear models are most readily specified through a model for the expected values and a model for the variances and covariances of the observables. After all, we are simply specifying (apart from the values of certain unknown parameters) the first two moments of our observables. Had their approach been in these terms, I believe it might still be discussed. Nelder (1977) had a related objection when he pointed out that randomization models (involving finite populations but random effects) could not be seen as a special case of the approach in [8]. The matter of providing linear unbiased estimates of quantities of interest figured nowhere in [8], and I believe this reduces many people's willingness to see its solution as general and relevant to their use of linear models and ANOVA. But perhaps the real reason that the description in [8] is not yet agreed upon is this: the majority of statisticians these days (perhaps even 50 years ago) are not interested in the issues that concerned Tukey, Cornfield, Kempthorne, Fairfield Smith and Neyman and co-workers, before them, and Nelder and others, including me, after them. Perhaps it is just too hard, connecting assumptions and models to the subject matter, to the data collection process, to the questions one is asking and the kinds of answers one seeks. "Does it really matter? Does it make any practical difference?" I get asked. It is so much easier discussing models and parameterizations.

5. Other ANOVA papers by Tukey.

5.1. *Dyadic ANOVA* [1]. This paper was based on a talk Tukey gave in November 1946, and is more interesting for what it tells us about the development of his thinking concerning ANOVA than for the material related to its title. Ostensibly about ANOVA for vectors, that is, what we would now call multivariate analysis of variance (MANOVA), the paper also contains a wealth of interesting material only marginally related to that topic. The reason he wrote it, he says, was that other accounts of MANOVA concentrate too much on tests and too little on that which is most useful and revealing in ordinary ANOVA. It is impossible to resist passing on one of his introductory remarks, presumably aimed at the average reader of *Human Biology*. He writes:

> It is a maxim of arithmetic that it is not proper to add 2 oranges to 1 apple; this is good arithmetic but may be poor vector algebra. For
>
> $$(2\,\text{oranges}, 0) + (0, 1\,\text{apple}) = (2\,\text{oranges}, 1\,\text{apple})$$
>
> is a meaningful and useful statement.

Later, he goes on:

> If we are to have an analysis of variance, we must have squares, and the solution is
>
> $$(2\,\text{oranges}, 1\,\text{apple})^2 = \begin{bmatrix} 4\,\text{orange}^2 & 2\,(\text{orange})(\text{apple}) \\ 2\,(\text{orange})(\text{apple}) & 1\,\text{apple}^2 \end{bmatrix}.$$

The paper includes a concise discussion of components of variance, initially in the context of Eisenhart's (1947) models, but also including the finite population pigeonhole models which were to play such a big role in his later work. Rather surprisingly in view of his later disdain for F-tests, and his stated motivation for writing the paper, he makes a start on tests of significance for dyadic ANOVA, that is, the distribution of eigenvalues in 2×2 MANOVA. He even attempts to give fiducial intervals for quantities of interest, but concludes that more distribution theory is required.

A topic not obviously related to dyadic ANOVA is what he calls choice of terms, that is, choice of the response variable to be analyzed in a given experiment. He castigates Fisher for not paying more attention to this point, illustrating it dramatically by carrying out the same analysis on some hydrogen spectrum data using both wavelength and its reciprocal, wave number, as responses. In a fascinating analysis foreshadowing the power transformation underpinnings of ODOFFNA, he uses his newly developed dyadic ANOVA to find that linear combination of a response variate and the variate squared which minimizes the ratio of row plus column sums of squares to interaction sum of squares in an unreplicated row-by-column array. Illustrating the method on one of the data sets which he uses in his later paper on ODOFFNA, Tukey shows the considerable gain in efficiency he achieves with his transformation. The eigenvalue problem he solves is reminiscent of canonical variate analysis, and he ends that discussion with some interesting speculations on alternative criteria to optimize in the definition of discriminant functions.

A further point of interest in this paper can be found in the Appendix, headed "Two identities and a lemma." The lemma gives the variance of the average and the expectation of the sample variance of a set of variates which have different means and different variances, but a common covariance λ, a simple enough variant on the result which is well known for i.i.d. variates. He goes on to apply this result to his pigeonhole models, illustrating once more what was to be a recurrent theme in his statistical research: a desire to weaken standard assumptions wherever possible. He finds that, under these more general assumptions, the formulae are essentially unchanged, with a common variance σ^2 being replaced by the average variance $\sigma_\cdot^2 - \lambda$.

5.2. *Components in regression* [4]. This paper is about simple linear regression when both variates are subject to "error," and the use of instrumental variates in this context. The fields of application discussed include precision of measurement, psychology and econometrics, and, as is so often the case with Tukey, the paper demonstrates the prodigious breadth of his knowledge. The connection with ANOVA is slight, really only arising because he discusses an example in which measurements are taken in replicate. As he says, "We could have avoided mention of variance components . . . since we only deal with the simplest sorts . . . between-vs-within or regression-vs-balance. However, we have chosen to bring them in for two reasons. Mainly to set the analysis in terms which can easily be carried over to more complicated analyses where the correct procedure might otherwise be a mystery. Secondarily, to stress the analogy with variance components for a single variate." The paper is not easy reading and, since its connection to other material here is not great, we do not discuss it any further.

5.3. *ANOVA and spectral analysis* [12]. As might be expected from its context—the discussion of two papers on the spectral analysis of time series—[12] is much more about spectral analysis than ANOVA. It was placed in one of the time series volumes [21], not in [17], yet I want to mention it here, in part for its influence on me personally. What Tukey makes very clear in this discussion is that spectrum analysis, with a *line* for each frequency, *is* ANOVA. More fully, he says "the spectrum analysis of a single time series is just a branch of variance component analysis." This was one of his inspired connections which proved illuminating in both directions. It is clear from his remarks that Tukey supposed that his statistical audience knew something about ANOVA and could read [8] if they wished, and that this would enlarge their understanding of spectrum analysis, the topic of the papers. What was probably not apparent at the time was that there were people, myself included, for whom spectrum analysis was straightforward, but variance component analysis a mystery, and that his connection would be helpful to such people in the other direction. For evidence of the impact of this paper on me, see Speed (1987); for a valuable introduction to this paper, see the comments by Brillinger in [21].

5.4. *Toward robust ANOVA* [16]. This paper offers "a recipe for robust/ resistant analysis of variance of data from factorial experiments in which all factors have three or more versions." Its motivation is eloquently explained as follows:

> Analysis of variance continues to be one of the most widely used statistical methods. Not only the form of the analysis of variance table with its lines of mean squares and degrees-of-freedom associated with each of several sorts of variation, but the entire analysis, including confidence statements, is classically supposed to be determined by the design—the hierarchical structure, conduct, and the intent of the experiment—alone. The behaviour of the data itself is, classically, not supposed to influence how its description is formatted. Hardly an exploratory attitude. . . . In this account, rather than using a data-free structure to define our procedure, we provide a further stage of responding to the data's behaviour, one where summarization is based on a robust alternative to the mean.

The recipe is explained by its application to a particular $5 \times 3 \times 8$ array of data from an experiment concerning the hardness of gold alloy fillings. It begins with a *pre-decomposition*, this being a multiway analogue of median polish, and proceeds through the *identification* of so-called exotic entries, to a *re-decomposition* dealing with these, and a *robust analysis of variance* with the familiar sums of squares and degrees-of-freedom calculated from the re-decomposition. Next, a process of *downsweeping* is carried out, this being a variant of the pooling of mean squares which we met in Section 3 above, and the recipe concludes with the calculation of error mean squares, standard errors and confidence statements.

5.5. *Methods, comments, challenges* [18–20]. Tukey expounded and discussed ANOVA in a number of his many overview papers, and I will single out three of these for brief mention.

In [18] he goes over "some methods that form sort of a general core of the statistical techniques" that were used at that time. He aimed "to supply background: statistical, algebraic and perhaps intuitive," and he succeeded admirably. The exposition could hardly be improved upon, indeed is better than most we see today, in that it contains possibly the first instance of the "analysis of variance diagram" mentioned in the discussion of paper [11] in Section 3 above. This diagram surely deserves to be more widely used. Also noteworthy is a remark which may well be the first appearance in print of the abbreviation ANOVA.

In [19] Tukey offers 37 methodological comments about statistics on topics ranging from *exploration versus confirmation*, *re-expression* and *causation*, to *spectrum analysis*, and naturally he has something so say about ANOVA. Relevant comments concern regression and analysis of variance, nonorthogonal analysis and

MANOVA, and can only be described as stimulating and provocative. For example, in seeking a replacement of conventional MANOVA: "We could calculate principal components, but they are not likely to be simply interpretable. So let us not"; and: "Much the same could be said of 'dust bowl empiricists factor analysis'."

In Section 21 of the last of these three overview papers [20], Tukey foreshadows the issues dealt with more fully in [16] discussed above. We see clearly how keen Tukey was to unify his understanding of and approach to ANOVA with his robust/resistant and exploratory data analysis paradigms. While [16] is a fine start, it seems clear that there is much more to be said on this unification.

6. Concluding remarks. John Tukey was an extraordinarily able and creative statistician. He made a number of lasting contributions to ANOVA: to our understanding of what it is and what it can do for us; to the algebraic and computational aspects of the subject; and, perhaps most important and characteristic to showing us how to go beyond the usual assumptions. The impact of all this work on the subject today is less than it should be, perhaps in part because Tukey set his standards rather high. However, his papers are all there for anyone to read, and if this appreciation of them encourages one person who would not otherwise, to do so, its purpose will have been achieved.

Acknowledgments. I thank Gordon Smyth, for reading this manuscript in draft and offering many helpful comments, and David Brillinger, for many valuable references and for his enthusiasm for this project. Thanks also to the referees and editors for their comments and assistance.

JOHN W. TUKEY'S PUBLICATIONS ON ANOVA

[1] (1949a). Dyadic ANOVA. An analysis of variance for vectors. *Human Biology* **21** 65–110. (Paper [21].)

[2] (1949b). One degree of freedom for non-additivity. *Biometrics* **5** 232–242. (Paper [29].)

[3] (1950). Some sampling simplified. *J. Amer. Statist. Assoc.* **45** 501–519. (Paper [38].)

[4] (1951). Components in regression. *Biometrics* **7** 33–69. (Paper [41].)

[5] (1955). Answer to query 113. *Biometrics* **11** 111–113. (Paper [51].)

[6] (1956a). Keeping moment-like sampling computations simple. *Ann. Math. Statist.* **27** 37–54. (Paper [52].)

[7] (1956b). Variances of variance components. I. Balanced designs. *Ann. Math. Statist.* **27** 722–736. (Paper [53].)

[8] (1956c). Average values of mean squares in factorials. *Ann. Math. Statist.* **27** 907–949. (With J. Cornfield. Paper [55].)

[9] (1957a). Variances of variance components. II. The unbalanced single classification. *Ann. Math. Statist.* **28** 43–56. (Paper [57].)

[10] (1957b). Variances of variance components. III. Third moments in a balanced single classification. *Ann. Math. Statist.* **28** 378–384. (Paper [58].)

[11] (1960). Complex analyses of variance: General problems. *Psychometrika* **25** 127–152. (With B. F. Green, Jr. Paper [76].)

[12] (1961). Discussion, emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics* **3** 191–219. (Paper [79].)

[13] (1962). Handout for Meeting of Experimental Station Statisticians "Tests for Non-additivity." ([17], Paper 3).

[14] (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA. (Book [307].)

[15] (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data. In *Design, Data and Analysis by Some Friends of Cuthbert Daniel* (C. Mallows, ed.) 171–244. Wiley, New York. (With E. G. Johnson. Paper [282].)

[16] (2001). Towards robust analysis of variance. In *Data Analysis from Statistical Foundations* (A. K. Md. E. Saleh, ed.) 217–244. Nova Science Publishers, Huntington, NY. (With A. H. Seheult. Paper [213].)

[17] (1992). *The Collected Works of John W. Tukey VII. Factorial and ANOVA*: *1949–1962*. Wadsworth, Belmont, CA.

The following are not entirely about ANOVA, but contain relevant material.

[18] (1951). Standard methods of analyzing data. In *Proc. Computation Seminar, December 1949* 95–112. IBM, Armonk, NY. (Paper [216].)

[19] (1980). Methodological comments focused on opportunities. In *Multivariate Techniques in Human Communication Research* (P. R. Monge and J. Cappella, eds.) 489–528. Academic Press, New York. (Paper [194].)

[20] (1982). An overview of techniques of data analysis, emphasizing its exploratory aspects. In *Some Recent Advances in Statistics* (J. Tiago de Oliveira and B. Epstein, eds.) 111–172. Academic Press, New York. (With C. L. Mallows. Paper [273].)

[21] (1984). *The Collected Works of John W. Tukey I. Time Series*: *1949–1964*. Wadsworth, Belmont, CA.

REFERENCES

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252.

DRESSEL, P. L. (1940). Statistical semi-invariants and their estimates with particular emphasis on their relation to algebraic invariants. *Ann. Math. Statist.* **11** 33–57.

EISENHART, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* **3** 1–21.

FISHER, R. A. (1929). Moments and product moments of sampling distributions. *Proc. London Math. Soc.* **30** 199–238.

HABERMAN, S. J. (1975). Direct products and linear models for complete factorial tables. *Ann. Statist.* **3** 314–333.

HARVILLE, D. A. (1969). Variances of variance-component estimators for the unbalanced 2-way cross classification with application to balanced incomplete block designs. *Ann. Math. Statist.* **40** 408–416.

KAPLAN, E. L. (1952). Tensor notation and the sampling cumulants of k-statistics. *Biometrika* **39** 319–323.

KEMPTHORNE, O. (1952). *Design and Analysis of Experiments*. Wiley, New York.

KENDALL, M. G. (1943). *The Advanced Theory of Statistics* **1**. Griffin, London.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, London.

MCCULLAGH, P. and PREGIBON, D. (1987). k-statistics and dispersion effects in regression. *Ann. Statist.* **15** 202–219.

NELDER, J. A. (1965a). The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proc. Roy. Soc. London Ser. A* **283** 147–162.

NELDER, J. A. (1965b). The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proc. Roy. Soc. London Ser. A* **283** 163–178.

NELDER, J. A. (1977). A reformulation of linear models (with discussion). *J. Roy. Statist. Soc. Ser. A* **140** 48–76.

NEYMAN, J. S. (1925). Contributions to the theory of small samples drawn from a finite population. *Biometrika* **17** 472–479. [Reprinted from *La Revue Mensuelle de Statistique* **6** (1923) 1–29.]

PAULL, A. E. (1950). On a preliminary test for pooling mean squares in the analysis of variance. *Ann. Math. Statist.* **21** 539–556.

PLACKETT, R. L. (1960). Models in the analysis of variance (with discussion). *J. Roy. Statist. Soc. Ser. B* **22** 195–217.

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.

SCHWARZ, C. J. (1993). The mixed-model ANOVA: The truth, the computer packages, the books. Part I. Balanced data. *Amer. Statist.* **47** 48–59.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York.

SEBER, G. A. F. (1977). *Linear Regression Analysis*. Wiley, New York.

SPEED, T. P. (1983). Cumulants and partition lattices. *Austral. J. Statist.* **25** 378–388.

SPEED, T. P. (1985). Dispersion models for factorial experiments. *Bull. Inst. Internat. Statist.* **51** 1–16.

SPEED, T. P. (1986a). Cumulants and partition lattices. II. Generalized k-statistics. *J. Austral. Math. Soc. Ser. A* **40** 34–53.

SPEED, T. P. (1986b). Cumulants and partition lattices. III. Multiply-indexed arrays. *J. Austral. Math. Soc. Ser. A* **40** 161–182.

SPEED, T. P. (1986c). Cumulants and partition lattices. IV. A.s. convergence of generalized k-statistics. *J. Austral. Math. Soc. Ser. A* **41** 79–94.

SPEED, T. P. (1987). What is an analysis of variance? (with discussion). *Ann. Statist.* **15** 885–941.

SPEED, T. P. and BAILEY, R. A. (1987). Factorial dispersion models. *Internat. Statist. Rev.* **55** 261–277.

SPEED, T. P. and SILCOCK, H. L. (1988a). Cumulants and partition lattices. V. Calculating generalized k-statistics. *J. Austral. Math. Soc. Ser. A* **44** 171–196.

SPEED, T. P. and SILCOCK, H. L. (1988b). Cumulants and partition lattices. VI. Variances and covariances of mean squares. *J. Austral. Math. Soc. Ser. A* **44** 362–388.

TRACY, D. S. and GUPTA, B. C. (1973). Multiple products of polykays using ordered partitions. *Ann. Statist.* **1** 913–923.

VOSS, D. T. (1999). Resolving the mixed models controversy. *Amer. Statist.* **53** 352–356.

WILK, M. B. and KEMPTHORNE, O. (1955). Fixed, mixed, and random models. *J. Amer. Statist. Assoc.* **50** 1144–1167.

WISHART, J. (1952a). Moment coefficients of the k-statistics in samples from a finite population. *Biometrika* **39** 1–13.

WISHART, J. (1952b). The combinatorial development of the cumulants of k-statistics. *Trabajos Estadistica* **3** 13–26.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720-3860
E-MAIL: terry@bilbo.berkeley.edu